Deficiencies in Current Practices of Clinical Natural Language Processing (CNLP) – White Paper

"Virtually *no* CNLP software is fit for purpose out-of-the-box and will invariably require tuning, if not significant enhancement, to serve a useful productive purpose *to a high accuracy for a particular client*."

Jon Patrick 2019

We have recently assessed the accuracy of Clinical NLP software available through either open source projects or commercial demonstration systems at processing pathology reports. This whitepaper discusses the twenty-eight deficiencies we observed in our testing of five different systems.

Our analysis is based on the need for industrial strength language engineering that must cope with a greater variety of real-world problems than that experienced by research solutions. In a research setting, users can tailor their data and pre-processing solutions to address the answer to a very specific investigation question unlike real-world usage where there is little, or no, control over input data. As a simple example, in a language engineering application the data could be delivered in a standard messaging format, say HL7, that has to be processed no matter what vagaries it embodies. In a research project that data could be curated to overcome the uncertainties created by this delivery mechanism by removing the HL7 components before the CNLP processing was invoked, a fix not available in a standard clinical setting.

When an organisation is intending to apply a CNLP system to their data the topics discussed in this document need to be assessed for their potential impact on their desired outcomes.

The evaluations were based on two key principles:

- There is a primary function to be performed by CNLP, that is, Clinical Entity Recognition (CER).
- There is one secondary function and that is Relationship Identification.

Any other clinical NLP processing will rely on one or both of these primary functions. For the purposes of this conversation we exclude "text mining" which uses a "bag-of-words" approach to language analysis and is woefully inadequate in a clinical setting.

Assessed Software:

- Amazon Concept Medical
- Stanford NLP + Metathesaurus
- OPenNLP + Metathesaurus
- GATE + Metathesaurus
- cTAKES

The systems have the listed deficiencies to a greater or lesser extent. No system has all these problems. The deficiencies discussed are compiled across the 5 systems under the following headings:

- Deficiencies in Understanding Document Structure
- Deficiencies in Tokenisation
- Deficiencies in Grammatical Understanding
- Deficiencies in the CER Algorithms
- Deficiencies in Semantics and Interpreting Medical Terminology

Deficiencies in Understanding Document Structure

1.

1.1 Missing Contextual Recognition

The first task for any system is to recognise the context of the text. This requires identifying the class of information in the document as a whole although sometimes it is only manifest through the structure of the document,

1.2 Inability to Recognise Headings.

Headings can be presented in a report by visual layout of uppercase or title case orthography and surrounding whitespace. However they can also be provided by lables from a HL7 tagset.

Headings provide key information on the shift in the type of content to be expected and therefore warrant a different processing objective, that is, key information components that represent major topic shifts. These are classically defined by section headers, but not always. Recognising headerless topic shifts is crucial to high accuracy results. Failure to recognise headings will lead to identification of incorrect entity values or inhibit corroboration of correct entity identification, e.g. identifying the full specimen description under examination might only be achieved by comparing content in the Final Diagnosis and the Nature of Specimen sections of a pathology report.

One system had difficulty recognising headings that where concatenations of words embedded with full stops, e.g

Pathology.Report.Section due to their tokenizers behaviour. As headings are important both for section boundary detection and context setting this Deficiency threatens a great deal of later processing.

Inability to properly recognise specimen boundaries.

1.3 Separating specimens in a multi-specimen report is critical to correct interpretation of the disease location. In some types of reports many specimens may be described with only some containing disease so incorrect identification of the boundary of the specimen description will result in the wrong specimen being assigned the identified disease.

Deficiencies in Tokenisation

2. Weaknesses in tokenisation

2.1 Tokens can be crudely defined as the strings between whitespace and they take many forms. A large range of non-alphabetic keyboard characters can be used for different purposes and in clinical texts the slash "/" has many functions. It can used to express a ratio but also to signify time duration, date, a proportion of lymph nodes involved in a malignant tumour, etc. Two tokenisers keep the tokens on each side of the slash together while another separated them, so that each was correct some of the time and incorrect at other times. This problem needs stronger context identification to produce the correct analysis at a consistently high accuracy level.

2.2

Deficiency to recognise alphanumeric entities

2.3 Many entities are described with a combination of characters and digits, especially biochemical names. These can be written with and without hyphens, e.g. HER2 and HER-2. It is not uncommon to see the numeric component treated incorrectly as the value of the entity in question instead of being part of its name.

Inability to exclude bullet point markers from any named entity

It is common to present content as a series of bullet points to make for easier reading. The bullet identifier can be of many different forms including digits, Roman digits in upper and lower case, dots and hyphens. Incorrect tokenisation has incorporated this information into a clinical entity, so that subsequently the entity could not be correctly semantically identified.

Faulty \Newline tokenising

We notice that different tokenisers use different ways to deal with the newline symbol '\n'. Three tokenisers do not split the input string by the newline symbol '\n'. Two Tokenisers separate the tokens by backslash '\'
2.4 and merge n into the next word. The third tokeniser keeps the whole newline symbol '\n' together as a whole token concatenated with the next word.

Faulty Interpretation of '\'

One tokeniser for a reason we don't entirely understand changed a '\' to a
2.5 '\\'. While their subsequent processing seemed to cope with this shift we found that all our annotations were made incorrect as to their position due to the introduction of new characters.

Faulty Special Symbol Tokenising

2.6 One of the tokenisers did not recognise these symbols, {'|','^','~'}, and treated them together with the neighbouring words giving faulty outcomes.

2.7 Problematic Interpretation of the Hyphen '-'

The use of the hyphen is ambiguous for CNLP. It can used to join two concepts together and to separate two discrete concepts from each other. Our policy is to separate lexical elements either side of a hyphen and interpret each individually. However, the three examined tokenisers do not split by hyphen and treat the whole combination as one single token.
2.8 This is bad practice.

Faulty Alphanumeric String Processing

Alphanumeric strings should usually be kept intact for clinical
2.9 processing as they most often represent a unit record identifier of some sort. One tokeniser split the string at character-type boundaries resulting in false identification.

Faulty tokenisation of real valued numbers

One tokeniser would split real numbers on the decimal point so as to create 3 tokens. This destroys the value of any decimal numeric values attached to attributes for example.

Deficiencies in Grammatical Understanding

3.

Missing Acronym Association with Expanded Name

3.1

- Lack of association of acronyms with their full names. Clinical reports are replete with acronyms and their accurate interpretation is important. Where they are presented along with their expanded name the two should be correctly attached to each whereas we have observed them being treated as separate entity references.

Context inconsistencies

3.2

- A weakness at identifying the same content in different contexts. Some systems are inconsistent in that they will identify a given entity correctly in one context but fail to identify the same entity in a different context. This is particularly surprising and indicates a lack of generalisation in their entity recognition function.

Inability to recognise the same entity with the same words expressed in a different word order.

3.3

A critical aspect of entity recognition is being able to recognise the same content with variable word order e.g. "high grade serous carcinoma" versus "Serous high grade carcinoma". Simple CER methods that use rule based approaches will have a serious difficulty with this common problem. Statistical machine learning methods are required to circumvent this by treating it as a generalised problem.

3.4

Failure to recognise Morphology of words

- We have observed an inability to identify entities when the same word is rendered in a different lexical morphology. Many words have the same general meaning but change form when used in different grammatical roles e.g. "malignant" and "malignancy". However at times the different morphology can also carry different meanings which often needs to be discriminated, so "malignant cells" is a description of a behaviour, whereas "malignancy" is a statement of a disorder.

3.5

Incomplete Negation Recognition

- Negation in clinical texts is of vital importance but is also complicated because of its four-way between the semantics of negative meaning and the grammar of negation representation, such as {normal, abnormal, not normal, not abnormal}. While some systems do recognise grammatical negation many do not control for the positive/negative aspect of the semantics of individual medical lexical items. This failure can lead to either false positives or false negatives in the processed outcome.

Part-of-Speech (POS) Identification Invalid

Two common mis-categorisations of POS is the assignment of nouns as adjectives and incorrect identification of Proper Nouns.

3.6 Erroneous Sentence boundary Detection

- Sentence Boundary detection was often faulty due to the tokenising of the newline character. When the "n character was concatenated with the following word often section names would become unidentifiable e.g. "nDiagnosis". We regard this as a major failure because of the cascading effect in correctly processing the document.

3.7

Deficiencies in the CER Algorithms

4. Inconsistent Relationship linking

- Identifying relationships between entities to a high accuracy is very difficult and still very much a research topic. Systems that do identify relationships need to be very careful but at least should be consistent in its pairings which, from our observations, commonly they are not.

4.1

4.2 Mistakes in Graphical Representation of Relationships

- Drawing lines to connect related entities and labeling graphically is helpful in interpreting the computational constructions, but they need to connect the correct entities, and not create false relationships. The danger here is that a visually appealing graphical representation carries a lot of weight and errors are therefore easily accepted.

4.3

Intrusive Newlines

- Interference in recognising the correct extent of an entity can be due to newline characters. Entity recognition can be seriously imperiled by newline characters distributed throughout the text as is very commonplace in pathology reports. It is important to use a pre-processor cleaning mechanisms to remove these extraneous characters so that an entity is properly recognised even if a non-printable character is buried within its extent.

5.

5.1

Deficiencies in Semantics and Understanding Medical Terminology

Unawareness of Anatomical Hierarchy

- Unawareness of conventional anatomical hierarchy. Some processing shows a lack of awareness of the general anatomical hierarchy, e.g. cell components mislabeled as anatomical class.

- Poor utilisation of body structure ontology.

Lack of a comprehensive medical vocabulary.

- Common medical words that should readily match to a known name go unrecognised, e.g. oophorectomy.

Non-discrimination of Meta-information

5.2 • Lack of discrimination between meta-information and patient specific information. Pathology reports always contain general information from the body of knowledge of the discipline and

5.3 specific content describing the patient's disease state e.g. descriptions of the criteria for selecting various values for a grade value, are not the grade of the sample examined.

Inability to identify Attribute-Value Pairs

- Inability to identify the difference between the name of an attribute

5.4 and a description of the patient's actual condition. Synoptic reports are designed to lay out the pertinent case information in a structure of Attribute-Value pairs. The attribute names do not represent an identified characteristic of the patient's health but rather are a label that can only be interpreted in conjunction with the value e.g. Lymphovascular invasion: Absent, is NOT a statement about the presence of this type of tumour invasion. Some systems ignore the structure and use the mention of the condition as validation that it exists for the patient. This of course leads to incorrect output.

5.5

Mis-labelling Unknown strings

- Automatically labeling any string it can't recognise as a Test Name entity. In any NLP system the processing of unrecognizable words needs to be very robust. We have observed the assignment of

5.6 sematic categories to these items which is both dangerous and needless in most cases.

Ambiguity in acronyms cannot be resolved correctly.

Acronyms, whilst highly useful in their own context, create confusion when they have alternative interpretations, e.g. MM can be either millimetres or Malignant Melanoma. Without proper identification of the context using statistical processing the correct interpretation cannot be made.